Lansdall-Welfare, T., Sudhahar, S., Thompson, J., & Cristianini, N. (2017). The Actors of History: Narrative Network Analysis Reveals the Institutions of Power in British Society Between 1800-1950. In *Advances in Intelligent Data Analysis XVI: 16th International Symposium, IDA 2017, London, UK, October 26–28, 2017, Proceedings* (Vol. 10584, pp. 186-197). (Information Systems and Applications, incl. Internet/Web, and HCI). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-319-68765-0_16

Peer reviewed version

Link to published version (if available):
10.1007/978-3-319-68765-0_16

Link to publication record in Explore Bristol Research
PDF-document

The Actors of History: Narrative Network Analysis Reveals the Institutions of Power in British Society Between 1800-1950

Thomas Lansdall-Welfare[1], Saatviga Sudhahar[1], James Thompson[2], and Nello Cristianini[1]

[1] Intelligent Systems Laboratory, University of Bristol, Bristol, United Kingdom
[2] Department of History, University of Bristol, Bristol, United Kingdom

Abstract. In this study we analyze a corpus of 35.9 million articles from local British newspapers published between 1800 and 1950, investigating the changing role played by key actors in public life. This involves the role of institutions (such as the Church or Parliament) and individual actors (such as the Monarch). The analysis is performed by transforming the corpus into a narrative network, whose nodes are actors, whose links are actions, and whose communities represent tightly interacting parts of society. We observe how the relative importance of these communities evolves over time, as well as the centrality of various actors. All this provides an automated way to analyze how different actors and institutions shaped public discourse over a time span of 150 years. We discover the role of the Church, Monarchy, Local Government, and the peculiarities of the separation of powers in the United Kingdom. The combination of AI algorithms with tools from the computational social sciences and data-science, is a promising way to address the many open questions of Digital Humanities.

Keywords: big data, network analysis, digital humanities, narrative analysis, natural language processing

1 Introduction

Previous successes in the application of Artificial Intelligence (AI) techniques to data analysis date back many decades [15], and have enabled massive progress in fields such as Bioinformatics, Physics and Social Sciences. This intelligent data analysis (IDA) has only recently started to be advocated as part of historical research in the burgeoning fields of the Digital Humanities, mostly fueled by the ongoing digitization efforts that involve many libraries around the world. Early attempts [19] have been met with skepticism by part of the historical community, whose main criticism could be summarized as the need for digital humanities to go beyond counting words [14].

Here, we look at the different roles played in British society by religion, monarchy and the various components of governance (judiciary, legislature, central and local government) and how they have changed over time, a change that

has been extensively studied using traditional methodologies (see, for example [16]). The boundaries between the various sources of legal authority (legislature, executive and judiciary) have not always been clear, and have been the object of discussion among scholars, particularly the separation between executive and legislative power, in the 19th century [2, 6]. Furthermore, if we focus on "soft power" (influence), then the boundaries between the sphere of action of legal and other moral authorities are even more indistinct, with the role of the Church and the Monarchy being crucial, yet ever shifting over the centuries.

The relation between these political power structures and their representation in language is studied in Critical Discourse Analysis [10], a discipline interested in what our use of language can reveal about political power relations in society. One way in which this can be done is to determine who controls and shapes the narrative - or discourse - of the news.

We are interested in adding a data-driven element to this discussion about the evolution of the spheres of power and influence in British society, by analyzing the discourse found in historical local newspapers published over 150 years. In particular, we are interested in charting the various spheres of action of different narrative communities in the overall narrative network of local-news content, and following the evolution of their boundaries and key actors.

As such, we present here the first application of large-scale Narrative Network Analysis [27] to a corpus of 35.9 million articles, made possible by the deployment of a dependency parser in a map-reduce framework, and the study of topological properties of the resulting narrative networks, a method built on the pioneering work of [12] and automated by the use of Natural Language Processing by [26]. The resulting analysis involves 29 networks made up of a total of 156,738 nodes and 230,879 edges connecting them extracted from 150 years of newspapers. Different spheres of inter-actions (entire sectors of society and governance), as well as the role of individual actors, are mapped by analyzing these narrative networks extracted from vast amounts of text. We are interested in the changes and continuities, in the role of individual actors as well as in that of entire communities.

Note that, in accordance with Critical Discourse Analysis, this method goes beyond the mere political and legal balance of powers, and includes the moral authority and influence on the collective imaginary exercised by authorities such as the Church and the Monarchy. Furthermore, it can include intellectuals, opinion leaders or commercial players (though this might be the subject of analysis for later corpora). In other words, we want to see how public discourse - and the narration of reality - are organized and segmented in the period 1800 to 1950.

2 Data Description

The newspaper collection used in this study is composed of 35.9 million newspaper articles from a combination of the British Newspaper Archive (FindMyPast, 2017) and digitized newspaper records provided by the Joint Information Systems Committee (JISC) from the same geographic regions and time period.

The collection was curated from the full archive so that the selected subset would allow for a comprehensive data-driven study of Britain in the 19th century on a representative sample of newspapers. The selection was performed by committee, and the criteria for selection included: the completeness of the digitization of a given newspaper title, the number of years that a newspaper title covers, the geographical region that the newspaper is from, the quality of the OCR output for the newspaper title and the political bias of the newspaper.

Our aim was to represent all geographical regions and time intervals as fairly as allowed by the available data. Newspaper titles were first split into their different geographical regions, and then within each region newspaper titles were ranked by a combination of the years covered (favoring titles with many years of continuous coverage), their average OCR quality, and the total size of data available for the title. Newspaper titles were then selected from this ranking until each geographical region had good coverage. We further used domain knowledge to take into consideration the balance of political opinion in the regional press at the time.

The newspaper collection we assembled, first reported on in [18], includes 28.6 billion words from 120 titles selected to best cover the United Kingdom, covering approximately 14% of the total output of U.K. regional newspapers during the period.

3 Methodology

Our methodology begins with the extraction of semantic triplets (subject-verb-object triplets where subject and object are noun phrases, referred to as 'actors', and the verb is transitive) from the corpus. The semantic triplets are used to generate a network of actors, linked by the transitive verbs connecting them in narrative. Analysis of the topology of these networks is used to discover important narrative information about the corpus [12, 26].

3.1 Extracting Semantic Triplets and Networks

Events are actions performed by actors that can be summed up by a verb or a name of an action [12]. An event can therefore be thought of as a narrative or story, where someone does something (the action) on or to someone or something else. Linguistically, these events can be represented as sequential sets of semantic triplets. Here, we consider semantic triplets of the form Subject-Verb-Object (SVO) which consists of a subject S as an actor, the verb V as the action performed by S, and O as the target of the action [11] (e.g., Police (S) Arrest (V) Thief (O)).

Before extracting semantic triplets, we first pre-process the textual data with a co-reference and anaphora resolution procedure. Co-reference resolution identifies whether two actors in the text refer to the same entity in the world [25]. These actors are then resolved to the most common short representation for the actor, using the Orthomatcher module in the ANNIE plugin of GATE, which

reports an average precision of 96% and recall of 93% [5]. Anaphora resolution is used to resolve pronouns in the text to the specific actor that is being referred to, and is performed using the Pronominal resolution module in ANNIE, which reports an average precision of 66% and recall of 46% [5].

After pre-processing, the resulting text is split into sentences and passed into a dependency parser [22], outputting the dependency tree of each sentence. From the dependency tree, we extract the sentence subject, the sentence verb, and the object of the verb into the appropriate form for the semantic triplets. In total, 140,225,349 semantic triplets were extracted from the newspaper content.

From these semantic triplets, we generate semantic networks where the nodes of the network represent subjects and objects and the edges represent the verbs connecting them. We prune the triplet networks to reduce noise, keeping only nodes (actors) which occur a minimum of three times in the extracted triplets.

Networks were generated for each decade covered by the corpus, with an overlap of five years, giving us a total of 29 networks. Overlapping of networks was chosen to allow us to extract the communities which persist throughout the time period under investigation.

3.2 Centrality of Nodes

The centrality of a node in a network can be measured in many ways, and can be used as one measure of the importance of a node to the connected structure of the network. Here we consider the betweenness centrality of nodes, which represents the number of shortest paths between nodes in the network that pass through a given node [13].

3.3 Community Detection

Community detection is the task of partitioning a network into different communities, where densely connected nodes are placed within the same community, with nodes belonging to different communities being sparsely connected. Within the semantic networks presented here, community detection therefore corresponds with finding the communities of actors within the text that often perform actions on or to each other, with sparser interaction between different communities of actors. Communities in our networks were discovered by computing the modularity class of each node within the Gephi software [3], which attempts to find communities based upon finding high modularity partitions of the network using Blondel's fast unfolding algorithm, which has been shown to outperform other community detection algorithms [4].

Once the communities for each node within each of the 29 networks had been computed, we linked the communities across time, allowing us to form "macro-communities" which represent persistent community structures that are exhibited in the majority of the different networks across the 150 year period.

For each actor, we built a vector of their network communities, where each element in the vector refers to the actors community in each year, resulting in a

Table 1: List of the most salient actors in the narrative networks found in the 150 years of historical British newspaper corpus using the two different measures of salience: frequency of an actor and the centrality of an actor.

Most Frequent Actors	One, This, Place, The, Mr, Defendant, Men, Prisoner, Members, Man, People, Nothing, Government, Committee, Bill, House, Some, More, Council, Attention
Most Central Actors	Bills, King, House, Government, God, Queen, Chairman, Committee, Duke, Court, England, London, President, Council, Bench, Prince, Jury, Clerk, Port, Men

29-length vector encapsulating all the communities in which that actor participates. The similarity between each node and every other node was then computed using the cosine similarity, generating a similarity matrix of the communities for each actor. Using this similarity matrix, a 'macro-network' of actors is generated, with an edge joining every two actors that have a cosine similarity greater than 0.5, corresponding to the two actors participating in the same community more often than not.

On the 'macro-network', we once again perform community detection using the same algorithm as before [4], partitioning the network into smaller partitions with high modularity. The resulting communities from this process are referred to as 'macro-communities' and represent the persistent community structures over time.

3.4 Sentimental Classification of Actions

Actions in the narrative can fall into one of three sentimental categories: positive, negative or neutral. We can classify each of the actions being performed in the narrative into one of these categories to determine how sentimental the portrayal of the actions performed by or to the actors of a given community are reported on average in the news. We use the linguistic resource SentiWordNet [1], containing the positive and negative polarity for a set of 13,767 verbs. For each verb extracted as part of a semantic triplet, we score the sentimental polarity of the action as the average difference in positive and negative polarity for the verb across the possible verb synsets. Sentimental polarity for a given community or macro-community is then calculated as the average polarity of all actions performed by or to that community respectively.

4 Results

In this study, we investigate the narrative structure of the news over 150 years by network analysis of actors and the actions that connect them. Each network we generated contains all narrative triplets in a 10 year period that occur at

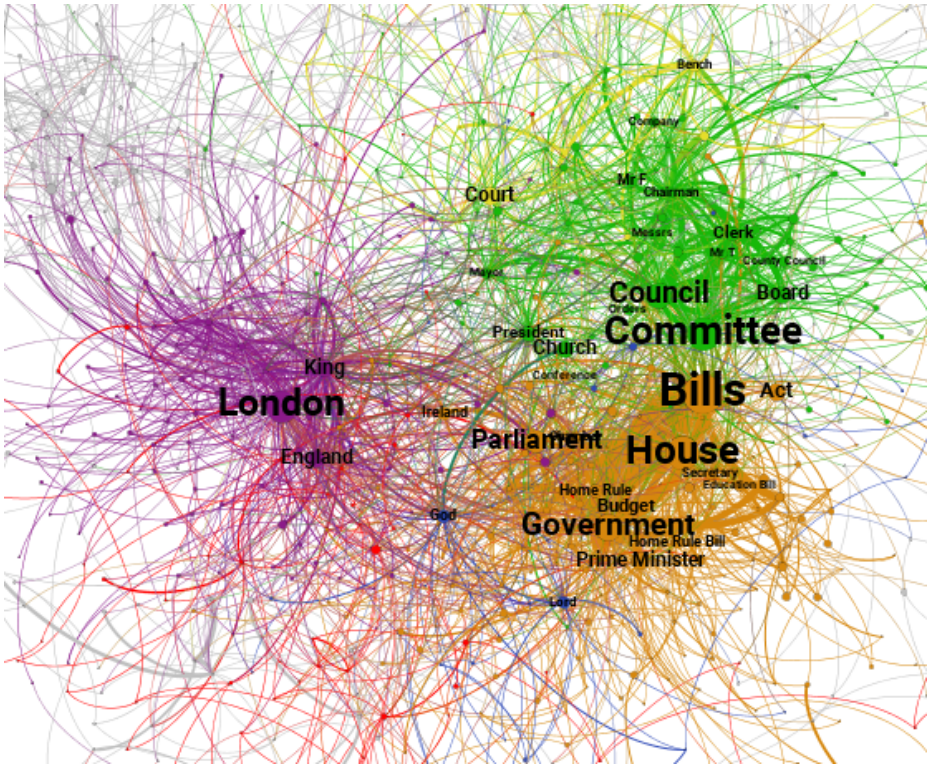


Fig. 1: Narrative network of the actors (nodes) and actions (edges) performed by them between 1905 and 1915 in the British newspaper corpus. Nodes are coloured based upon the community to which they belong.

least three times, with a five year overlap, giving us a total of 29 networks in all. An example[3] of one of these networks for the triplets extracted between 1905 and 1915 can be seen in Fig. 1. These narrative networks are formed by actors (entities who act or are acted upon) linked by their actions. In these networks, there are two possible notions of salience for an actor: its raw frequency within the corpus or its centrality within the network. The second notion, that of centrality, reflects how closely embedded this actor is into the overall structure, or how well linked it is to all other actors. It is a measure of how close all actors are to this actor. Using these two measures of salience for an actor, Table 1 shows an overview of the 20 most frequent actors, along with the 20 most central actors, extracted from the news corpus.

We found that simply using the frequency of an actor does not generate a very meaningful list of actors, with "this", "one", "the", "defendant", "men" being

[3] The full list of networks, along with their properties can be seen at `http://thinkbig.enm.bris.ac.uk/supp-info-actors-of-history/`

those most frequent in all the triplets. However, we discover that the most-central actor in the narrative is "Bills", speaking to the centrality of politics in the coverage of local and national events, and in shaping public discourse in Britain. Following this, we discovered the next most central actors are "King", "House", "Government", "God", and "Queen", additionally showing the influence of both the monarchy and religion in the image of society communicated to the public via the mass media. This difference in meaning between notions of salience for an actor shows the importance of the linkage of an actor with the overall narrative.

Building up from this, communities in a narrative network are subsets of actors that are interacting tightly with each other in the narrative, and interacting much less with others, so that they can be separated from the rest of the network with a relatively small number of cuts. By identifying the key communities, we partition the set of all actors into groups of highly interacting players, that signal power (hard and soft) structures within public life. For example, one community found in the discourse contains the actors "Court", "Jury", "Bench", "Judge", "Magistrates", etc. which we shall label as the Judiciary community.

Through the analysis of the central actors, and of the communities found in the narrative network, we discover three clear findings, relating to the division of powers of the State, the significance of local government intervention, and role of influence and moral authority in society.

4.1 Division of Powers

We discovered that the macro-communities within the narrative networks clearly identify the main spheres of action within society: Judiciary, Church, Monarchy, Local Government - but that they do not separate between the legislature and executive (Central Government). This can be seen in Fig. 2 where we show the macro-communities discovered for the 1000 most central actors over the 150 year period, finding that the orange community (labeled as 'Central Government') of nodes contain actors such as "Government", "Prime Minister", "Home Secretary" and "Lord Chancellor", but also "Bills", "House", "Commons" and "Parliament". This is in line with a long-standing position among scholars [2] according to which the U.K. has in practice been lacking the separation between those two powers. The key actors for the other structures of society include "Council", "Chairman", "Board" and "Clerk" (labeled as 'Local Government'), "Judge", "Jury", "Court" and "Counsel" (labeled as 'Judiciary'), "King", "Queen", "Prince" and "Princess" (labeled as 'Royalty') and finally "Bishop", "Rev", "Lord Bishop" and "Archbishop" (labeled as 'Church').

4.2 Local Government

As one might expect, we find that a prominent role is played by the 'central' (e.g. 'London'-centered) actors, such as "House", "Commons" and 'Prime Minister", but we also see, particularly from the 1870s onwards, the increasing accessibility and activism of Local Government over time. We can see that the actor "Board" gains in prominence after the Public Health legislation of 1848, but with

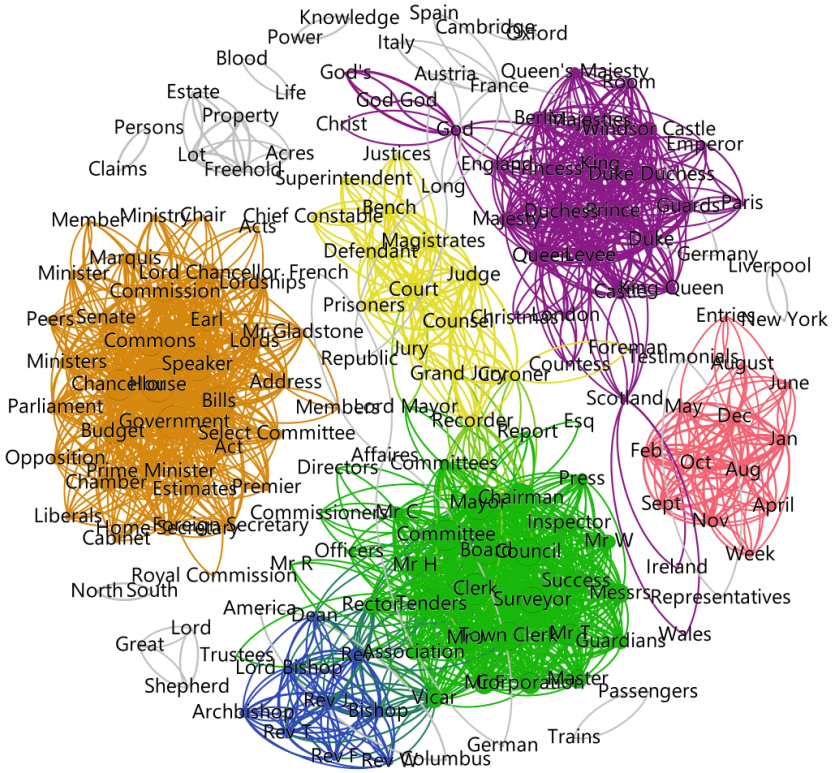


Fig. 2: Macro-communities discovered for the 1000 actors with the highest centrality over the 150 year period. It can be seen that the largest communities correspond with the broad topics of executive/legislative (orange, left), local authority (green, bottom right), royalty (purple, top right), judiciary (yellow, centre), months (pink, right) and the church (blue, bottom).

some added salience from roughly the 1870s, that reflects the creation of School Boards, elected bodies tasked with creating and providing elementary education for children in areas that were under-served by the existing schools (see Fig. 3). By the last two decades of the 19th century, a nexus appears to emerge around "Committee", "Council", "Chairman", "Clerk", sometimes "Mayor", persisting into the mid-20th century. This growth in the relevance of Local Government can be seen in the change between the 1830s (Fig. 4a) when the main communities in the narrative are Central Government and Royalty and the 1900s (Fig. 4b) where we can clearly see that Local Government has become a major player in the narrative of the media.

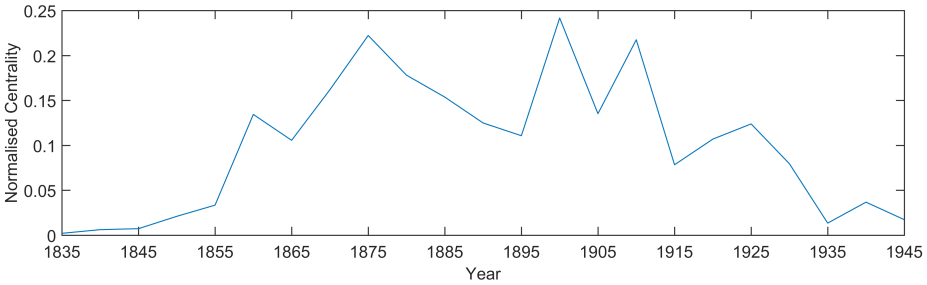


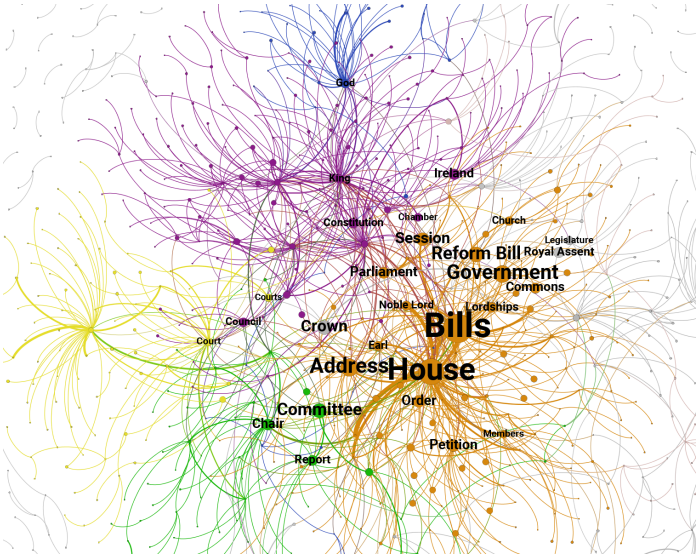
Fig. 3: Normalised centrality of the actor "Board" between 1830 and 1950 showing the prominence of the actor "Board" after the Public Health legislation of 1848.

4.3 The Role of Influence and Moral Authority

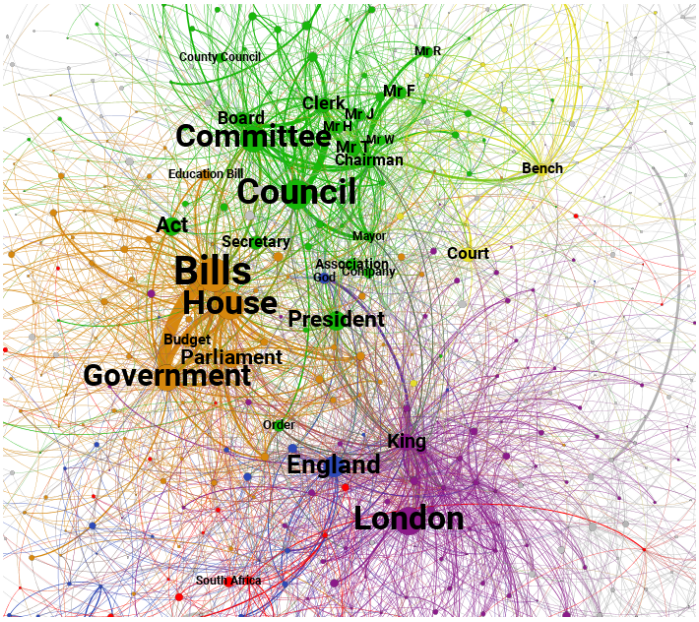
While the actual power of the Monarchy is known to have been declining during the period under investigation [8, 17], we see that its role, along with that of the other members of royalty, in public discourse is very prominent throughout the period, manifesting as a distinct macro-community in the narrative network, populated by actors such as "King", "Queen", "Majesty", "Prince", and "Princess". From the late 19th century the Monarchy is known to have made a deliberate effort to maintain a presence in the media [17]. The centrality of the actors in the community - especially that of the monarchs – speaks to the cultural prestige and popularity that the Crown retained even as its actual power was limited. This is further supported by an analysis of the sentimental classification of the actions (§ 3.4), where we found that, on average, the Royalty community is portrayed as having the most positive actions performed on them. This coverage of royalty does surely bring out the differences between narrative appeal and political power, as the latter for the monarchy is known to be much less by 1950 than it had been in 1800 [9].

A similar case can be seen in the role of the Anglican community, formed by actors such as "Lord Bishop", "Archbishop", "Bishop", and "Rev". The Church of England as the established church held long-term significance to local life of the people, but its role in society was declining over the years, slowly becoming less central in the narrative of our society as presented in the news, supporting the view presented in [7], where it is argued that the 1960s are the decade of real change for the Church in which it loses considerable cultural salience. Notice that we do see, as expected in the case of the United Kingdom, a separation between Church and Government.

We also found that another centre of power in a State, the military, appears to form a large community during times of war but is not a constant presence throughout the period, and so does not appear in the macro-communities.



(a) 1830-1840



(b) 1900-1910

Fig. 4: The growth in the narrative role played by Local Government can be seen in (a) and (b), showing in (a) that the community as a whole, centred around the "Committee" and "Chair" nodes (shown in green, bottom left) did not play a major role in the 1830s, but, as shown in (b), by the start of the 20th century was an integral part of society (centred around the nodes "Committee" and "Council", top).

5 Conclusion

This study is aimed at addressing the criticism leveled at digital humanities approaches to historical corpora, which had called for efforts to go "beyond counting words". We show that the deployment of AI techniques to this data-analysis task allows us to access valuable semantic information that would not be accessible without Intelligent Data Analysis.

We also observe that this approach represents a step towards the full automation of the literary approach that Franco Moretti called distant reading [20]: the combination of AI, big data and data-visualization methods to large corpora has the potential to turn massive textual resources into semantically meaningful representations. This has an enormous potential in a community where most standard methods still involve various forms of statistical counting on words [19, 21, 23, 24].

Acknowledgments

Thomas Lansdall-Welfare, Saatviga Sudhahar and Nello Cristianini are supported by the ERC Advanced Grant "ThinkBig" awarded to NC. The authors would like to thank FindMyPast for making the original corpus available for study, as well as Dr. Gaetano Dato for his helpful comments.

6 References

1. Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.
2. Walter Bagehot. *The English Constitution*, volume 3. Kegan Paul, Trench, Trübner, 1900.
3. Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks, 2009.
4. Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
5. Kalina Bontcheva, Marin Dimitrov, Diana Maynard, Valentin Tablan, and Hamish Cunningham. Shallow methods for named entity coreference resolution. In *Chaines de références et résolveurs danaphores, workshop TALN*, 2002.
6. Anthony Wilfred Bradley and Keith D Ewing. *Constitutional and administrative law*, volume 1. Pearson Education, 2007.
7. Callum G Brown. *The death of Christian Britain: understanding secularisation, 1800–2000*. Routledge, 2009.
8. David Cannadine. The context, performance and meaning of ritual: the british monarchy and the invention of tradition, c. 1820–1977. In *The Invention of Tradition*, pages 101–64. Cambridge University Press Cambridge, 1983.
9. David M Craig. The crowned republic? monarchy and anti-monarchy in britain, 1760–1901. *The Historical Journal*, 46(01):167–185, 2003.

10. Norman Fairclough. *Critical discourse analysis: The critical study of language*. Routledge, 2013.
11. Roberto Franzosi. Narrative as data: Linguistic and statistical tools for the quantitative study of historical events. *International review of social history*, 43(S6):81–104, 1998.
12. Roberto Franzosi. *Quantitative narrative analysis*. Number 162. Sage, 2010.
13. Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.
14. Paul Gooding. Mass digitization and the garbage dump: The conflicting needs of quantitative and qualitative methods. *Literary and Linguistic Computing*, 28(3):425–431, 2013.
15. David J Hand. Intelligent data analysis: Issues and opportunities. In *Advances in Intelligent Data Analysis. Reasoning about Data: Second International Symposium, IDA-97, London, UK, August 1997. Proceedings*, page 1. Springer, 1997.
16. Jose Harris. *The Penguin Social History of Britain: Private Lives, Public Spirit: Britain 1870-1914*. Penguin UK, 1994.
17. Eric Hobsbawm. The invention of tradition edited by eric hobsbawm and terence ranger, 1983.
18. Thomas Lansdall-Welfare, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper Team, and Nello Cristianini. Content analysis of 150 years of british periodicals. *Proceedings of the National Academy of Sciences*, page 201606380, 2017.
19. Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182, 2011.
20. Franco Moretti. *Distant reading*. Verso Books, 2013.
21. Bob Nicholson. Counting culture; or, how to read victorian newspapers from a distance. *Journal of Victorian Culture*, 17(2):238–246, 2012.
22. Joakim Nivre, Johan Hall, and Jens Nilsson. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, volume 6, pages 2216–2219, 2006.
23. Eitan Adam Pechenick, Christopher M Danforth, and Peter Sheridan Dodds. Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one*, 10(10):e0137041, 2015.
24. Steffen Roth, Carlton Clark, and Jan Berkel. The fashionable functions reloaded: An updated google ngram view of trends in functional differentiation (1800-2000). 2016.
25. Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544, 2001.
26. Saatviga Sudhahar. *Automated Analysis of Narrative Text using Network Analysis in Large Corpora*. PhD thesis, University of Bristol, 2015.
27. Saatviga Sudhahar, Gianluca De Fazio, Roberto Franzosi, and Nello Cristianini. Network analysis of narrative content in large corpora. *Natural Language Engineering*, 21(01):81–112, 2015.